Evolutionary Plasticity of Protein Families: Coupling Between Sequence and Structure Variation

Anna R. Panchenko,[1]* Yuri I. Wolf,[1] Larisa A. Panchenko,[2] and Thomas Madej[1]
[1]*Computational Biology Branch, National Center for Biotechnology Information, National Institutes of Health, Bethesda, Maryland*
[2]*Department of Biology, Moscow State University, Moscow, Russia*

ABSTRACT **In this work we examine how protein structural changes are coupled with sequence variation in the course of evolution of a family of homologs. The sequence–structure correlation analysis performed on 81 homologous protein families shows that the majority of them exhibit statistically significant linear correlation between the measures of sequence and structural similarity. We observed, however, that there are cases where structural variability cannot be mainly explained by sequence variation, such as protein families with a number of disulfide bonds. To understand whether structures from different families and/or folds evolve in the same manner, we compared the degrees of structural change per unit of sequence change ("the evolutionary plasticity of structure") between those families with a significant linear correlation. Using rigorous statistical procedures we find that, with a few exceptions, evolutionary plasticity does not show a statistically significant difference between protein families. Similar sequence–structure analysis performed for protein loop regions shows that evolutionary plasticity of loop regions is greater than for the protein core. Proteins 2005;61:535–544.**

Key words: protein structural evolution; sequence variation; protein loops; sequence–structure correlation

INTRODUCTION

A protein sequence folds into a unique, highly ordered conformation which maintains its specific function. As proteins evolve, their sequences change due to amino acid replacements, the majority of which are believed to be effectively neutral.[1] Consequently, protein-specific function, structure, folding, and the protein–protein interaction network as a rule change gradually in the course of evolution. Indeed, the overall protein structural topology is so well preserved throughout evolution that proteins that diverged billions of years ago may still show remarkable structural resemblance and, in many cases, sequence conservation as well.[2]

The fundamental question of whether protein structures evolve by divergence or by convergence inspired many comparative studies of protein structures and networks of protein similarities.[3–10,42] According to the convergent scenario, protein structural similarity can occur independently in two proteins due to the limited number of topological arrangements.[11,12] Recently, it has been shown that convergent models do not adequately describe the patterns of sequence and structural similarity observed in the populations of real proteins by using graph theoretical methods.[8,10] By contrast, the scale-free behavior and other important characteristic features of protein networks can be correctly reproduced using divergent models of structural evolution.[7–10] In these models, new protein structures emerge, and existing structures change through the processes of duplication and subsequent divergence from a common ancestor.

The sequence and structural analysis of many commonly observed protein folds points to the dominant role of divergent mechanisms in protein structural evolution as well.[13–17] It has been demonstrated, for example, that proteins from the TIM barrel, OB-fold, cupredoxin, and β-trefoil folds have common features in their topology, nature of ligands, and location of catalytic residues, which points to the plausibility of divergent scenarios for these and other protein folds comprising the protein universe. In a previous study, we likewise observed a significant linear correlation between sequence similarity and loop structural similarity for the aforementioned folds.[18] Given that the loops do not contribute much to the protein core stability, we argued that the strong coupling between the changes in sequence and loop structure can only happen due to divergent evolution.

Chothia and Lesk first addressed the question of coupling between the structural and sequence changes in proteins, and found an exponential dependence of root-mean-square deviation on percent of sequence identity.[2] Further studies that were performed on larger datasets of proteins showed similar results.[5,19] Recently, however, it has been shown on a sample of 36 protein families that most of the structural variation in aligned regions of homologous proteins is linearly correlated with the changes in sequence which supports the "global" model of protein

Grant sponsor: the NIH Intramural Research Program

*Correspondence to: Anna Panchenko, Computational Biology Branch, National Center for Biotechnology Information, Building 38A, National Institutes of Health, Bethesda, MD 20894. E-mail: panch@ncbi.nlm.nih.gov

Received 28 March 2005; Accepted 12 May 2005

Published online 23 September 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20644

structure.[20] According to this model, all residue–residue interactions, not just a few key residues, are important in determining the unique protein structure. In an attempt to solve the "fold recognition" problem and design structural models for new sequences, Koehl and Levitt performed an analysis of how structural changes between two protein folds correlate with the differences between the sequences that are compatible with these folds.[21] They also found, on a benchmark of 12 protein families, that structural changes as measured by cRMS are linearly related to the changes in sequence.

In this article we study how the protein structure changes in its conserved aligned core regions and unaligned loop regions as proteins diverge from a common ancestor. We performed a sequence–structure correlation analysis on a large number of families of homologous proteins and found a statistically significant linear correlation between measures of sequence and structural similarity for the great majority of these families. This finding allows us to address the next important question of how much sequence change can protein structure tolerate, and whether it depends on the type of protein fold, or on some other sequence and structural characteristics. We call this quantity "the evolutionary plasticity of structure" (EPS), and estimate it by calculating the regression coefficients of linear sequence–structure dependencies for homologs.

METHODS

Test Set

Sets of homologous protein families were extracted from the CDD search database version 1.62 at http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml. The CDD collection of protein domain alignments includes curated CDDs[22] and preprocessed domain families imported from SMART and PFAM, 6222 protein domain families altogether.[23] Upon import, the sequences from SMART/PFAM alignments with more than 75% identity with known structures were substituted by the most similar structure from the Protein Data Bank.[24] Those families containing short sequence repeats and having average alignment length of less than 50 residues were excluded from the test set.

Each CDD family was decomposed into a set of pairwise structure–structure alignments. Structural alignments within CDD families were computed by the VAST algorithm,[25] and were selected for analysis according to the following criteria: (a) the mutual overlap between the VAST alignment footprint and CDD footprint (the footprint for a given sequence was defined as a region between the first and the last residues aligned by VAST or CDD) was at least 80%; (b) X-ray resolution of both structures in a pair was better than 3.0 Å; (c) BLAST E-value calculated for VAST alignment was less than 0.01; (d) any discontinuous domain[26] inconsistently aligned between VAST and CDD was disregarded.

Additionally, to the requirements imposed on structural pairs we selected protein families based on the following criteria: (a) the protein family should contain at least 10 structurally aligned protein pairs; (b) proteins from a given family should span a wide range of sequence similarity, that is, should cover a range of at least 30% in sequence identity between the most diverged and least diverged structural pair; (c) not more than two protein family alignments from the same domain cluster were retained in the final test set; the redundancy between protein families was checked by using the procedure implemented in the CDART algorithm.[27] Even though these protein families can belong to the same domain cluster, they are coming from different sources and have rather different alignments (Table I).

The final test set comprised 81 CDD families covering a wide range of functional and structural classes. The list of test families together with their length, number of protein pairs, and the PDB code of the first structure is shown in Table I. The test set for loop analysis contained 59 families, excluding 22 families that had a high fraction of pairs with missing coordinates in loop regions (see the next section).

Measures of Structural and Sequence Similarity

To measure the quality of linear correlation between sequence and structural characteristics for homologous proteins from the same family, we first need to choose the most sensitive and reliable measures of sequence and structural similarity. Because most of the structural similarity measures (RMSD, AHM, LHM) are extensive and depend on the number of residues and protein size, the aforementioned structural measures should be divided by the radius of gyration (similar but not identical results were obtained with the normalization by the square root of the number of aligned residues). The radius of gyration for a protein pair was calculated for each of the two proteins in the pair based on the structurally aligned part and then was averaged. As a result, the normalized RMSD, AHM, and LHM quantities do not depend on the number of residues any more. Nonnormalized conventional measures of structural similarity yielded weaker sequence/structure correlation (not shown) so that in our further analysis we used only normalized structural similarity measures.

The sequence similarity was measured as the BLAST bitscore[28] divided by the alignment length (bitscore per residue). Structural similarity measures based on comparing the structures in the aligned regions comprised RMSD, fraction of conserved contacts (CC), and aligned Hausdorff measure (AHM), whereas the loop-based Hausdorff measure (LHM) quantified the difference in the loop regions. The fraction of conserved contacts was calculated as a fraction of identical residue contacts in both structures divided by the average number of contacts in both structures made by the aligned residues.[29] The contacts were defined between residues separated along the chain by at least five peptide bonds and having C^{α} atoms less than 8 Å apart.

The root-mean-squared deviation (RMSD) was calculated using the superposition algorithm due to McLachlan.[30] Another measure that quantified the structural difference of proteins between the aligned regions and between the loops was based on the mathematical concept of Hausdorff distance.[18,31] Let $A = \{a_1, \ldots, a_m\}$ and

TABLE I

Identifier[a]	PDB[b]	Len[c]	N[d]	Corr (ρ)[e]	Slope (b)[f]	Description[g]
cd00157, smart00174	1I4D_D, 1M7B_A	172,173	66, 10	−0.21, −0.51	−0.017, −0.039	Rho subfamily of Ras-like small GTPases [P-loop containing nucleotide triphosphate hydrolases]
pfam00969	1JK8_B	86	32	−0.53	−0.066	Class II histocompatibility antigen, beta domain [MHC antigen-recognition domain]
smart00085, pfam00068	1BK9, 1BK9	102,102	210,102	−0.44, −0.39	−0.074, −0.081	Phospholipase A2 [Phospholipase A2, PLA2]
pfam00017	2SHP_A	86	21	−0.58	−0.078	SH2 domain [SH2-like]
cd00119, smart00263	1LMQ, 1C7P_A	109,116	24, 67	−0.55, −0.69	−0.088, −0.095	C-type lysozyme and alpha-lactalbumin [Lysozyme-like]
smart00651	1B34_B	63	30	−0.59	−0.09	snRNP Sm proteins [Sm-like]
cd00367	1POH	85	10	−0.84	−0.092	Histidine-containing phosphocarrier protein (HPr) [HPr-like]
pfam00077	1MVP_A	84	15	−0.77	−0.092	Retroviral aspartyl protease [Acid proteases]
smart00034, cd00037	1TN3, 1IOD_B	90, 93	35,263	−0.67, −0.61	−0.097, −0.101	C-type lectin (CTL) or carbohydrate-recognition domain (CRD) [C-type lectin-like]
smart00429	1NFI_B	97	21	−0.9	−0.098	Ig-like, plexins, transcription factors [Immunoglobulin-like beta-sandwich]
pfam00406	3AKY	174	28	−0.45	−0.104	Adenylate kinase [P-loop containing nucleotide triphosphate hydrolases]
pfam00030, smart00247*	1A45, 1ELP_A	81, 76	10, 15	−0.79, −0.91	−0.08, −0.107	Beta/gamma crystallins [gamma-Crystallin-like]
smart00125	21BI	130	10	−0.89	−0.112	Interleukin-1 homologs [beta-Trefoil]
cd00070	1QKQ_A	124	28	−0.8	−0.114	Galectin/galactose-binding lectin [Concanavalin A-like lectins/glucanases]
pfam00502	1QGW_C	148	15	−0.75	−0.114	Phycobilisome protein [Globin-like]
pfam00073*, cd00205*	1EAH_1, 1VBB_1	216,195	95, 71	−0.87, −0.85	−0.097, −0.114	Picomavirus capsid protein domain [Viral coat and capsid proteins]
pfam01833*	1BFS	89	39	−0.89	−0.119	IPT/TIG domain [Immunoglobulin-like beta-sandwich]
pfam00259*	1A0E_A	381	28	−0.99	−0.12	Xylose isomerase [TIM beta/alpha-barrel]
pfam00129*	1K8D_A	175	28	−0.96	−0.122	Class I Histocompatibility antigen, domains alpha 1 and 2 [MHC antigen-recognition domain]
pfam02800	1JN0_O	153	39	−0.79	−0.122	Glyceraldehyde 3-phosphate dehydrogenase, C-terminal domain [Glyceraldehyde-3-phosphate dehydrogenase-like, C-terminal domain]
smart00636*, pfam00704	1KFW_A, 1LG1_A	350,285	36, 53	−0.88, −0.83	−0.123, −0.187	Glycosyl hydrolases family 18 [TIM beta/alpha-barrel]
pfam00074, cd00163*	1K2A_A, 1QMT_A	98, 99	44, 25	−0.61, −0.89	−0.066, −0.123	Pancreatic ribonucleases [RNase A-like]
pfam00248*	1QRQ_C	277	28	−0.93	−0.124	Aldo/keto reductase family [TIM beta/alpha-barrel]
pfam00056*	1LLC	135	44	−0.83	−0.125	lactate/malate dehydrogenase, NAD binding domain [NAD(P)-binding Rossmann-fold domains]
cd00190, smart00020	1MKX_K, 1DLE_A	211,208	378,561	−0.57, −0.57	−0.126, −0.15	Trypsin-like serine protease [Trypsin-like serine proteases]
pfam00194*	1ZNC_A	245	10	−0.99	−0.127	Eukaryotic-type carbonic anhydrase [Carbonic anhydrase]
pfam00107*	1KOL_A	337	64	−0.93	−0.129	Zinc-binding dehydrogenase [GroES-like; NAD(P)-binding Rossmann-fold domains]
cd00148*	1D1J_D	120	15	−0.92	−0.13	Profilin [Profilin-like]
pfam00686*	4CGT	94	15	−0.93	−0.131	Starch binding domain [Prealbumin-like]
pfam00043*	1K3Y_A	107	77	−0.82	−0.134	Glutathione S-transferase, C-terminal domain [Glutathione S-transferase (GST), C-terminal domain]
pfam00061	1PMP_C	131	55	−0.62	−0.134	Lipocalin/cytosolic fatty-acid binding protein family [Lipocalins]
cd00099*	1CDI	105	133	−0.84	−0.134	Immunoglobulin domain variable region (v) subfamily [Immunoglobulin-like beta-sandwich]
pfam00112*, smart00645*	1EF7_A, 3PBH	200,202	55, 90	−0.84, −0.87	−0.137, −0.15	Papain family cysteine protease [Cysteine proteinases]
pfam00101*	1IR2_7	102	15	−0.91	−0.137	Ribulose bisphosphate carboxylase, small chain [RuBisCO, small subunit]
smart00631*	1H8L_A	260	15	−0.91	−0.138	Zn_pept domain [Phosphorylase/hydrolase-like]
cd00098	2HRP_L	88	85	−0.69	−0.14	Immunoglobulin domain constant region 1 (cl) subfamily [Immunoglobulin-like beta-sandwich]
pfam00394*	1GS6_X	118	16	−0.97	−0.149	Multicopper oxidase [Cupredoxin-like]
pfam00210*	1KRQ_A	152	19	−0.92	−0.149	Ferritin-like domain [Ferritin-like]
pfam00016*	1RUS_A	236	10	−0.92	−0.15	Ribulose bisphosphate carboxylase large chain, catalytic domain [TIM beta/alpha-barrel]

TABLE I. Continued

Identifier[a]	PDB[b]	Len[c]	N[d]	Corr (ρ)[e]	Slope (b)[f]	Description[g]
pfam00144*	1BLH	264	45	−0.91	−0.151	Beta-lactamase [beta-Lactamase/D-ala carboxypeptidase]
cd00195, smart00212	2UCZ, 2UCZ	140, 141	21,45	−0.64, −0.8	−0.155, −0.157	Ubiquitin-conjugating enzyme E2 and UBC homologues [UBC-like]
pfam00155*	1B8G_B	353	14	−0.93	−0.157	Aminotransferase class I and II [PLP-dependent transferases]
smart00102	1AHQ	116	10	−0.89	−0.161	Actin depolymerisation factor/cofilin-like domains [Gelsolin-like]
pfam00135	1QO9_A	485	28	−0.8	−0.161	Carboxylesterase [alpha/beta-Hydrolases]
cd00051	1FW4_A	57	59	−0.76	−0.162	EF-hand, calcium binding motif [EF Hand-like]
smart00235*, cd00203*	1CIZ_A, 1HV5_A	137, 134	34,23	−0.88, −0.92	−0.163, −0.206	Zinc-dependent metalloprotease [Zincin-like]
pfam00258*	1CZU_A	143	26	−0.89	−0.167	Flavodoxin [Flavodoxin-like]
pfam02866*	1LDN_H	143	29	−0.88	−0.167	lactate/malate dehydrogenase, alpha/beta C-terminal [Lactate and malate dehydrogenases, C-terminal]
pfam00331*, smart00633*	1HIZ_A, 1HIZ_A	297, 297	15,21	−0.95, −0.93	−0.168, −0.185	Glycosyl hydrolase family 10 [TIM beta/alpha-barrel]
smart00452*, pfam00197*	1AVU, 1BA7_A	151, 150	10,10	−0.92, −0.9	−0.171, −0.185	Soybean trypsin inhibitor (Kunitz) family of protease inhibitors [beta-Trefoil]
pfam00141*	1ITK_B	240	48	−0.93	−0.174	Peroxidase [Heme-dependent peroxidases]
pfam00161	1PAG_A	232	28	−0.87	−0.174	Ribosome inactivating protein [Ribosome inactivating proteins (RIP)]
pfam00111, cd00207*	1L5P_A, 1AWD	69, 78	73,38	−0.78, −0.88	−0.165, −0.176	2Fe-2S iron-sulfur cluster binding domain [beta-Grasp (ubiquitin-like)]
cd00314*	1FHF_A	236	76	−0.91	−0.178	Plant peroxidase superfamily [Heme-dependent peroxidases]
pfam00180	1HQS_A	343	10	−0.93	−0.181	Isocitrate/isopropylmalate dehydrogenase [Isocitrate/Isopropylmalate dehydrogenases]
pfam02806*, smart00632*	1C8Q_A, 1KXQ_A	78, 81	39,31	−0.93, −0.94	−0.181, −0.188	Alpha amylase, C-terminal all-beta domain [alpha-Amylases, C-terminal beta-sheet domain]
cd00047*, smart00194*	1GWZ, 2SHP_A	228, 248	28,25	−0.89, −0.93	−0.186, −0.214	Protein tyrosine phosphatases (PTP), catalytic domain [(Phosphotyrosine protein) phosphatases II]
pfam00042	4VHB_A	133	96	−0.76	−0.186	Globin [Globin-like]
pfam00076	1A9N_B	72	15	−0.68	−0.191	RNA recognition motif [Ferredoxin-like]
pfam00127	1AIZ_A	81	87	−0.85	−0.196	Copper binding proteins, plastocyanin/azurin family [Cupredoxin-like]
pfam00227*	1JD2_F	189	56	−0.84	−0.21	Proteasome A-type and B-type [Ntn hydrolase-like]
pfam00337	1C1F_A	122	15	−0.89	−0.213	Galactoside-binding lectin [Concanavalin A-like lectins/glucanases]
pfam00080*	1F1G_C	139	15	−0.98	−0.218	Copper/zinc superoxide dismutase (SODC) [Immunoglobulin-like beta-sandwich]
pfam00208*	1EUZ_E	203	11	−0.94	−0.272	Glutamate/Leucine/Phenylalanine/Valine dehydrogenase [NAD(P)-binding Rossmann-fold domains]

[a]CDD profile identifier.
[b]Representative PDB structure.
[c]Domain alignment length.
[d]Number of structural pairs in a family.
[e]Linear correlation coefficient.
[f]Slope of the linear regression line (Evolutionary Plasticity of Structure).
[g]CDD domain description [SCOP 1.63[41] fold is listed in square brackets].
The table lists protein families together with the PDB code of the template structure, average length of pairwise structure–structure alignments, number of structural pairs per family, Pearson linear correlation coefficient between AHM and BLAST bitscore per residue, slope of the regression lines (regression coefficient or EPS), and the family description. Those families having high linear correlation (correlation coefficients obtained with both normalized RMSD and normalized AHM were less than −0.8 and r^2-ratio for both measures of structural similarity were higher than 0.9) are indicated by the asterisks.

$B = \{b_1, \ldots, b_n\}$ be finite point sets in a Euclidean space. The Hausdorff distance between the sets A and B is then defined by:

$$d_H(A,B) = \max\{\min_j d(a_1,b_j), \ldots, \min_j d(a_m,b_j), \min_i d(a_i,b_1), \ldots, \min_i d(a_i,b_n)\} \quad (1)$$

Here, the terms $d(a_i,b_j)$ denote the Euclidean distance between the points. In other words, the Hausdorff distance between the sets A and B is the smallest distance such that every point $a_i \in A$ is within this distance of some point $b_j \in B$, and vice versa. Hausdorff distance can be defined under the assumption that the structural alignment between two

domains is known and the C^{α} atoms for both structures are in a common coordinate frame.

The Hausdorff measure for loops (LHM) was calculated as follows:

$$LHM = \frac{1}{n_s - 1} \sum_{i=1}^{n_s - 1} h_i. \quad (2)$$

Here "loop" is defined as a region between two consecutive aligned secondary structure elements and n_s is the number of aligned secondary structure elements: $h_i = 0$, if the ith loop regions do not have any unaligned residues; $h_i = d_H(A_i, B_i)$, where A_i contains the set of C^{α} coordinates of nonaligned residues in the ith loop of the first structure in a pair, the last aligned residue from the preceding aligned region, and the first aligned residue from the following aligned region. Similarly, B_i is defined for the second structure in a pair. The sets (A_i, B_i) are defined to include two aligned residues so that the measure can be defined even if one of the sets of nonaligned residues is empty. In the calculation of LHM, those pairs where one or the other protein had more than 25% missing residues in nonaligned loops were excluded. In the case of AHM, instead of the coordinates for the C^{α} atoms in the loops, we use the coordinates for the C^{α} atoms in the aligned segments and average over the number of aligned segments.

Definitions of disulfide bonds were obtained from the PDB files of all protein structures for each family. Bonds formed outside of the structure–structure alignment footprint regions (see "Test set" section) were disregarded. The average number of disulfide bonds per family was calculated as the sum of the number of SS-bonds in each protein in a family divided by the number of proteins. The fraction of conserved disulfide bonds was calculated as a ratio between the number of identical SS-bonds in a protein pair and the average number of disulfide bonds within the footprint regions of two proteins.

Statistical Analysis

The statistical analyses described in this study used the Splus statistical package(version 6). To investigate the relationship between sequence and structural similarity we performed correlation and regression analyses. The Pearson linear correlation (ρ) and Spearman rank correlation coefficients were calculated, and the p-value under the null hypothesis that the correlation coefficient was equal to zero was estimated. Those families with p-values less than 0.01 were considered as having correlation coefficients significantly different from zero. To quantify how much the nonlinear terms improve the data fitting we included a quadratic term in the linear model and performed nonlinear regression analysis. The ratio of squared linear correlation coefficient for the linear model (R_l^2) and squared multiple correlation coefficient for the nonlinear model (R_n^2) ($r^2 = R = R_l^2/R_n^2$) in this case would indicate the relative improvement in the data fitting upon inclusion of the nonlinear term in the model. The higher this ratio is, the lower the contribution of nonlinear terms upon data fitting.

TABLE II

	Median correlation coefficient	Fraction of families with statistically significant correlation, %	Fraction of families with $r^2 > 0.9$, %
RMSD	−0.86	94	74
CC	0.74	84	77
AHM	−0.87	98	77
LHM	−0.75	88	71

The table shows the 50% quantile of Pearson correlation coefficients between AHM and BLAST bitscore per residue, fraction of families with statistically significant correlation (p-value less than 0.01) and the fraction of families with the ratio r^2 higher than 0.9 for each measure of structural similarity used in the study.

The F-test has been used to test the null hypothesis that all regression coefficients are equal, with alternative hypothesis being that the regression coefficients are not all equal. The null hypothesis has been rejected, and therefore we employed multiple comparison procedures. First we checked which regression coefficients were different from each other by using the Tukey-Kramer method.[32] For the purpose of illustrating the Tukey-Kramer method, the approximate method proposed by Gabriel can be applied, which computes the comparison intervals for all regression coefficients.[32] According to Gabriel's method, two regression coefficients are considered significantly different if and only if their comparison intervals do not overlap.

RESULTS

The Quality of Sequence–Structure Correlation for Different Protein Families

Table II shows the accuracy of correlation obtained between the BLAST bitscore per residue and various measures of structural similarity (RMSD, CC, AHM, and LHM). As can be seen from this table, the linear correlation is strong for most of the families, and half of them have correlation coefficients better than 0.73–0.87, depending on the structural similarity measure used (Table II lists Pearson correlation coefficients; Spearman rank correlation coefficients give similar results). This result is consistent with the studies of Wood and Pearson,[20] who showed on a smaller test set of 35 protein families that half of them have correlation coefficients greater than 0.878. Comparing different measures of structural similarity, one can see that normalized AHM tends to yield a stronger correlation than other quantities yielding 98% of families with statistically significant linear correlation coefficients (with p-value <0.01). In agreement with this observation, our previous studies showed that the AHM measure performs very well in distinguishing homologs from analogs.[18] High accuracy of the AHM is due to the higher sensitivity of the Hausdorff measure to subtle dissimilarities between the aligned parts of protein structures. Based on this observation, we chose this quantity to characterize the structural change in the present analysis.

Figure 1(a–d) illustrates the high quality of linear correlation for four protein families: Picornavirus capsid protein (pfam00073), Pancreatic ribonuclease (cd00163),

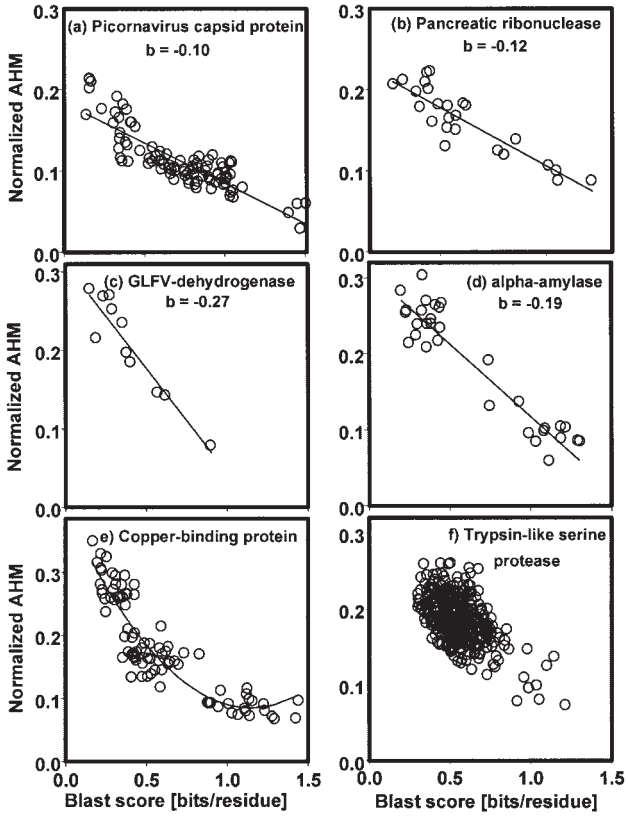

Fig. 1. Normalized AHM is plotted versus BLAST bitscore per residue for (**a**) Picornavirus capsid protein (pfam00073), (**b**) Pancreatic ribonuclease (cd00163), (**c**) GLFV-dehydrogenase (pfam00208), (**d**) Alpha-amylase (smart00632), (**e**) Copper binding proteins family (pfam00127), and (**f**) Trypsin-like serine protease (cd00190). Solid lines show the linear regression fit of the data and the values of the linear regression coefficient *b* are listed for the four top families. For the family of Copper binding proteins the solid line shows the nonlinear regression fit with quadratic term.

GLFV-dehydrogenase (pfam00208), and Alpha-amylase (smart00632), which all have Pearson linear correlation coefficients less than −0.87. As shown in Figure 1(e–f), not all families, however, exhibit such good correlation between sequence and structure changes. The Trypsin-like serine protease family (cd00190), for example, has a correlation coefficient of only −0.57 [Fig. 1(f)], while the Copper-binding proteins family (pfam00127) is more adequately described by the nonlinear regression model taking into account higher order quadratic terms (r^2-ratio being equal to 0.88) [Fig. 1(e)]. In the overall test set, among those with statistically significant correlation (79 families), 17 families had an r^2-ratio smaller than 0.9 indicating that, for these cases, adding the nonlinear term improves the performance of modeling by about 10%. It should be noted that alignments from different sources but belonging to the same protein family (see Methods, Table I) except for three cases exhibit consistent behavior with respect to the quality of linear correlation. Furthermore, random exclusion of duplicate families does not have any effect on the quality of linear correlation, nor on the results discussed below.

Although the correlation between protein sequence and structure is found to be statistically significant for the great majority of test families, there is still a high degree of variability in the magnitudes of the correlation coefficients among the families. There seems to be no strong relationship between the domain length (i.e., the average length of structure–structure alignments in a family) and the quality of linear correlation ($\rho = -0.30$, p-value = 0.01). No connection between correlation coefficients and contact density ($\rho = -0.23$, p-value = 0.04) or contact order[33] ($\rho = -0.27$, p-value = 0.02) has been observed either.

One might hypothesize that changes in structure should not always be strongly coupled with changes in amino acid sequence, especially if protein stability is determined mainly by the set of strong interactions such as covalent disulfide bonds. Figures 2 and 3 show how the quality of linear correlation depends on the disulfide bond content in protein families. As can be seen from Figure 2, protein families having on average two or more disulfide bonds per family (Sample 1, 13 families) exhibit rather poor sequence–structure correlation and proteins from the families with high correlation coefficients usually contain less than two disulfide bonds (Sample 2, 68 families). We should note that the difference between these two distributions is not caused by the difference in the family length (there is no significant correlation between the number of disulfide bonds per family and protein length).

To test the difference between two distributions of correlation coefficients (Sample 1 and Sample 2), we applied the Wilcoxon two-sample test, which showed that these two samples come from populations with different mean values (the null hypothesis was rejected with the p-value = 0.0016). We found that the majority of S—S bonds in Sample 1 were well conserved among different family representatives (more than 75% conserved S—S bonds) except for the three cases of Carboxylesterase (pfam00135, 72% conserved S—S bonds), Trypsin-like serine protease (smart00020, 71% conserved S—S bonds), and Papain family Cysteine protease (pfam00112, 63% conserved S—S bonds), whereas two of these families (pfam00135 and pfam00112) are also characterized by high sequence–structure correlation ($\rho = -0.80$, $\rho = -0.84$).

Figure 3 shows as well that the quality of sequence–structure correlation depends on the average number of disulfide bonds per family (the correlation coefficient is 0.44 with p-value of 0.001). Because not all disulfide bonds are conserved in protein families, we also calculated the fraction of conserved S—S bonds per family and showed in this figure those families that had the fraction of conserved S—S bonds higher than 0.5 (Fig. 3, crosses). A high fraction of conserved S—S bonds in a family points to the preservation of specific S—S bonds in evolution and can be used as a measure of reliability of their definition (correlation coefficient for data points shown by crosses is equal to 0.64 with p-value of 0.0007).

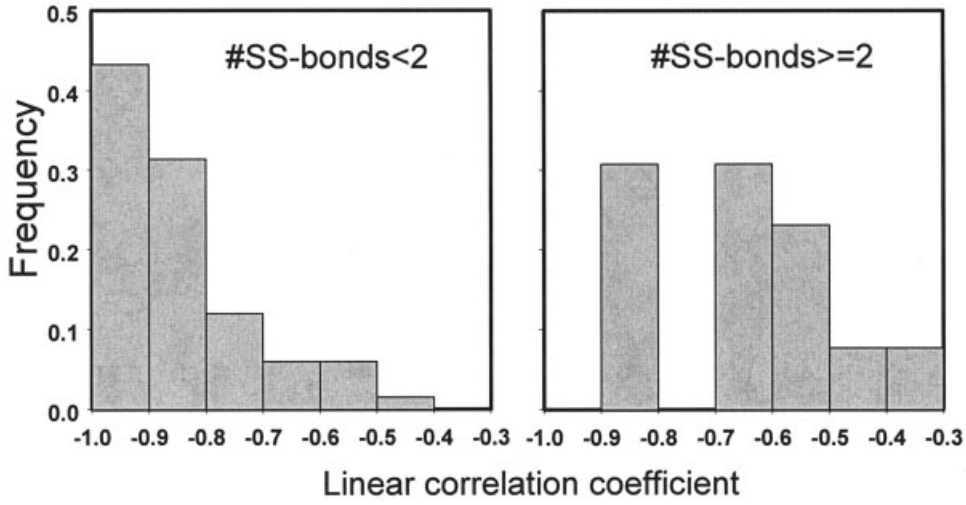

Fig. 2. The histogram shows the Person correlation coefficients between AHM and BLAST bitscore per residue for protein families with less (**a**) and more (**b**) than two disulfide bonds per family.

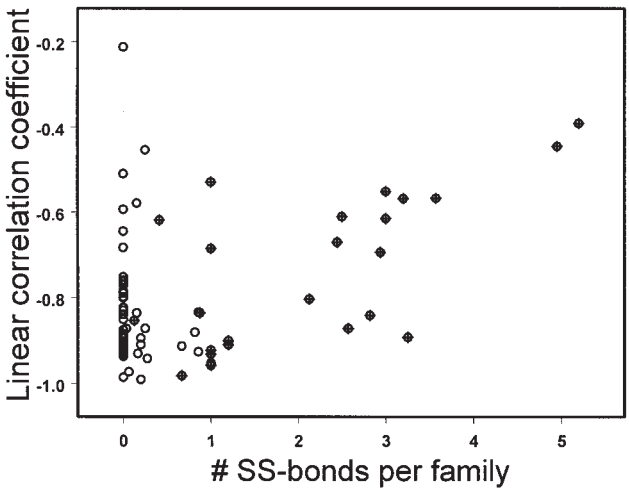

Fig. 3. Pearson correlation coefficient plotted against the number of disulfide bonds per family for the overall test set (circles) and only for those families which have more than 50% conserved disulfide bonds (crosses).

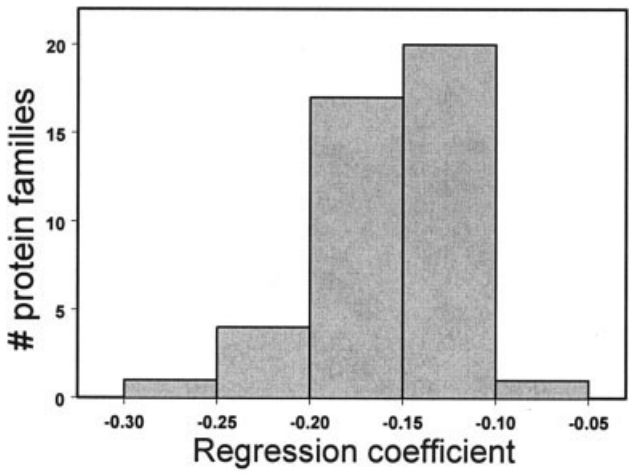

Fig. 4. The histogram shows linear regression coefficients for each family with high correlation (see the caption for Table I).

The Evolutionary Plasticity of Structure Estimated for Different Protein Families

As we showed in the previous section, for the majority of families, the sequence–structure dependence can be quite well described by the linear regression. The regression coefficients (the slope of the regression line) in these cases would estimate the relative structural to sequence change in the evolution of a particular protein family or, in other words, "the evolutionary plasticity of structure" (EPS). This measure is discussed below in more detail. To compare regression coefficients for different protein families, first we excluded families with poor correlation ($\rho_{RMSD} > -8.0$ or $\rho_{AHM} > -0.8$) and large contribution of nonlinear terms ($r^2_{RMSD} < 0.9$ or $r^2_{AHM} < 0.9$). This filtering procedure resulted in 43 families with high linear correlation (these families are marked by asterisks in Table I). Figure 4 depicts the histogram of regression coefficients for this set of 43 protein families. As can be seen from this figure, the EPS varies by about a factor of 3 among different protein families. Likewise, Wood and Pearson[20] reported a 3.9-fold change in their "structural mutation sensitivity" for a similar but smaller test set.

Although the regression coefficients vary between families, one needs to test whether this difference is statistically significant. To compare the slopes of the various families, we first tested the null hypothesis that all regression coefficients are equal (see Methods). This hypothesis is rejected with $P << 0.0001$. To determine which families have different structural tolerances, we employed multiple comparison methods and calculated the comparison intervals (95% confidence) for the regression coefficients of every protein family (Fig. 5). The comparison intervals are constructed such that two regression coefficients are significantly different if and only if their intervals do not overlap.[32] As can be seen from Figure 5, there

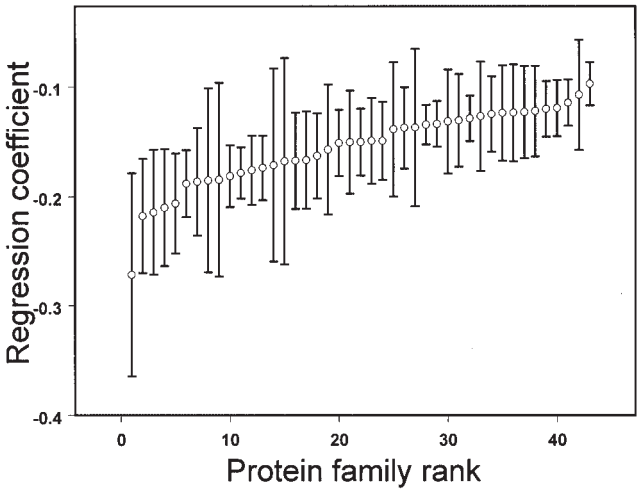

Fig. 5. The linear regression coefficients (*b*) are plotted together with their comparison intervals (see Methods) for each family with high correlation (see the caption for Table I). All families are ordered with respect to the increasing regression coefficients.

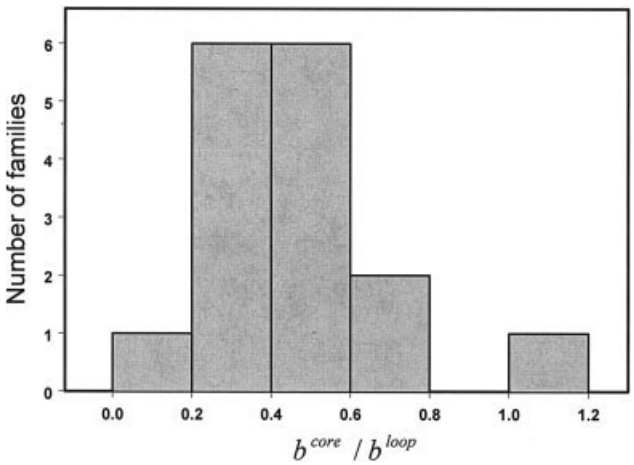

Fig. 6. The histogram of the ratio between regression coefficients obtained for aligned parts (AHM used as a measure of structural similarity) and regression coefficients obtained for loops (LHM used as a measure of structural similarity).

are apparently two groups of protein families that have significantly different regression coefficients and nonoverlapping comparison intervals, while the rest of the protein families do not exhibit a significant difference in slopes between each other.

The first group consists of several protein families having the steepest slopes (highest EPS) and positioned in the left side of the plot. These include GLFV-dehydrogenases (pfam00208, b = −0.27), Copper/zinc superoxide dismutase (pfam00080, b = −0.22), Protein tyrosine phosphatase (smart00194, b = −0.21), and Proteasome A-type and B-type (pfam00227, b = −0.21). The second group is formed by proteins with the smallest EPS, which are positioned on the right side of Figure 5; among them are Picornavirus capsid protein family (pfam00073, cd00205, b = −0.10), Beta/gamma-crystallins (smart00247, b = −0.11), IPT/TIG domain (pfam01833, b = −0.12), and Xylose isomerase (pfam00259, b = −0.12). Interestingly enough, some protein families characterized by the lowest EPS, form large interaction interfaces with other proteins or cell components. For example, Picornavirus capsid proteins are packed in highly ordered icosahedral shells that are maintained through multiple interactions between the subunits whereas crystallins, IPT/TIG and Xylose isomerase domains also participate in macromolecular interactions.

Overall, we found that EPS values for the majority of protein families do not differ significantly between each other because their comparison intervals (see Methods) overlap. Because our test protein families spanned a wide range of structural folds (Table I) and functions, the previous observation implies that EPS, in general, depends neither on the structural class nor on the protein fold type. For example, the Glycosyl hydrolase family (smart00633) has an EPS of −0.18, whereas the aldo/keto reductase/K+ channel beta subunit family has an EPS of about −0.12, although both protein families have the TIM barrel fold. The superfolds, the most populated structural topologies (TIM barrels, beta trefoils, four-helical bundles, and others), show EPS values comparable to those of other folds (not shown).

The Evolutionary Plasticity Is Different in Loop Regions Compared to the Protein Core Regions

The evolutionary relatedness between proteins can be successfully gauged from the comparison of their loop regions.[18,34] Table II shows that, within the families of homologous proteins, structural changes in loops are strongly coupled with the evolutionary distance which, in this case, was measured by the normalized BLAST bitscore for the aligned regions. The sequence–structure dependence in loop regions for 71% of protein families (the test set for the loop analysis, see Methods) can be well described by a linear model and, for 88% of the protein families the linear correlation coefficients are found to be statistically significant. Among families with a particularly high sequence–LHM correlation, are the families of Xylose isomerase, Class I Histocompatibility antigen, Protein tyrosine phosphatase, IG-like plexins, and others. For some families, for example, Ribonuclease A, the sequence–structure correlation for loops is even higher than the correlation observed for aligned core regions. The linear sequence–structure correlation suggests that loop regions are, in general, under constant evolutionary pressure, which preserves their overall structure and they therefore change gradually as proteins diverge.

To compare the EPS of aligned core regions with the EPS of loop regions, we computed the ratio of their regression coefficients ($b^{\mathrm{core}}/b^{\mathrm{loop}}$). The test set depicted in Figure 6 comprises 16 protein families with a good linear correlation for both LHM and AHM (with the requirement that both correlation coefficients are less than −0.8 and $r^2 > 0.9$). Assuming equal plasticity of core regions and loops (the null hypothesis), we expect that, in half of the instances, $b^{\mathrm{core}}/b^{\mathrm{loop}}$ ratios will fall below 1, and in half of

the instances these ratios will be above 1 (8:8 ratio). However, we observed 15 cases where the b^{core}/b^{loop} ratio was less than 1. The probability to observe such bias given the above assumption can be estimated from the binomial distribution as $p(0.5, 0, 16) + p(0.5, 1, 16) = 0.00026$. Thus, equal plasticities of core regions and loops is not likely to be compatible with our observations. This suggests that loop regions have higher evolutionary plasticity of structure compared to the protein core and, as can be seen from Figure 6, for the majority of families (12 families), the ratio of regression coefficients for the core and loop regions lies between 0.2 and 0.6.

DISCUSSION AND CONCLUSION

In this article, we study the structural evolution of homologous proteins in terms of their sequence–structure dependence. We showed that the protein structural variability for a great majority of protein families is linearly coupled with the sequence variability, which suggests that, typically, protein structure gradually changes as proteins diverge during evolution. However, when the protein structural core is stabilized by strong interactions such as disulfide bonds, the correlation between structural and sequence divergence is much weaker if detectable at all. Protein families that have large number of disulfide bonds (which are usually conserved) typically do not show a linear sequence–structure correlation in contrast to families with fewer disulfide bonds. Apparently, during the evolution of these families, purifying selection preserves the disulfide contacts and has a much weaker effect in the rest of the protein molecule such that, in these cases, the structural variability cannot be explained predominantly by the changes in sequence.

Drawing an analogy with solid mechanics, the sequence–structure dependence curves can be viewed as stress–strain curves where the physical body undergoes geometrical deformation after applying a stress. In the case of protein evolution, amino acid substitutions introduce the stress on protein structure, and structure either adjusts to the change or breaks apart. The linear dependences of measures of structural similarity on sequence similarity observed for the majority of protein families in our test set allows us to compare "the evolutionary plasticity of structure" (EPS) between different families. The evolutionary plasticity of structure for a given family is defined, accordingly, as a degree of structural variation per unit of sequence variation. Low values of EPS (shallow slope of the regression line) correspond to the situation when protein structure is highly conserved within a family of homologs relative to sequence changes. This could be caused either by strong functional constraints imposed on the structure or by high structural stiffness, that is, the inability to accommodate large structural variations without breaking the molecule apart. High values of EPS (steep slope) correspond to the situation when large structural shifts (within a framework of a given protein fold) can occur upon minor sequence divergence as a result of relaxed functional constraints on the structure and/or high structural tolerance of a given fold.

The rigorous statistical analysis performed in this work suggests that, with several exceptions, the values of the EPS for protein structural cores do not significantly differ between protein families. Interestingly enough, despite the variability among protein families in functional constraints and types of structural folds, the proteins from different families respond similarly to the sequence drift in evolution. This observation is based on the evaluation of multiple comparison intervals for the EPS values rather than on direct comparison of sequence–structure correlation slopes as has been done by others.[20] One could argue that this result could be an artifact caused by possible flaws in the analysis such as insufficient structural data and/or derivation of sequence and structure similarity measures. However, the observed high correlation between sequence and structural divergence within individual families suggests that the analysis described here is robust. Moreover, the observed EPS values were not found to be statistically different, even though the test set was designed in such a way (protein families with high linear correlation and sufficient number of sequences) to reduce the uncertainty of the EPS estimates.

It is commonly observed that the size of the sequence space is much larger than the size of structure space, and the number of different structural folds is rather small, estimated to be several thousand.[35–40] Moreover, certain protein topologies are realized in evolution much more often than others (so-called "superfolds"), and the existence of such inequality in fold frequencies is sometimes attributed to specific physicochemical or geometrical properties of superfolds. Our results demonstrate that the gradual change of structure follows the same pattern in different protein families, suggesting that the role of intrinsic characteristics of superfolds in evolution might be exaggerated. In this respect we argue that the differences between common and rare folds may arise in evolution semirandomly, that is, via self-enhancing stochastic fluctuations of abundance of essentially equal folds.[7] In any case, until the existence and significance of differences in "evolutionary plasticity of structure" between protein families is conclusively demonstrated, there is probably no ground to use their inequality as a working hypothesis in studies of protein structural evolution.

ACKNOWLEDGMENTS

We thank Stephen Bryant (NCBI), Eugene Koonin (NCBI), and Nick Grishin (University of Texas Southwestern Medical Center) for helpful discussions, and Lewis Geer for help with CDART database.

REFERENCES

1. Kimura M. The neutral theory of molecular evolution. Cambridge: Cambridge University Press; 1983.
2. Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. EMBO J 1986;5:823–826.
3. Holm L, Sander C. Mapping the protein universe. Science 1996;273: 595–603.
4. Murzin AG. How far divergent evolution goes in proteins. Curr Opin Struct Biol 1998;8:380–387.
5. Russell RB, Saqi MA, Sayle RA, Bates PA, Sternberg MJ. Recognition of analogous and homologous protein folds: analysis of

sequence and structure conservation. J Mol Biol 1997;269:423–439.
6. Matsuo Y, Bryant SH. identification of homologous core structures. Proteins 1999;35:70–79.
7. Koonin EV, Wolf YI, Karev GP. The structure of the protein universe and genome evolution. Nature 2002;420:218–223.
8. Dokholyan NV, Shakhnovich B, Shakhnovich EI. Expanding protein universe and its origin from the biological Big Bang. Proc Natl Acad Sci USA 2002;99:14132–14136.
9. Qian J, Luscombe NM, Gerstein M. Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. J Mol Biol 2001;313:673–681.
10. Deeds EJ, Shakhnovich B, Shakhnovich EI. Proteomic traces of speciation. J Mol Biol 2004;336:695–706.
11. Finkelstein AV, Ptitsyn OB. Why do globular proteins fit the limited set of folding patterns? Prog Biophys Mol Biol 1987;50:171–190.
12. Ptitsyn OB, Finkelstein AV. Similarities of protein topologies: evolutionary divergence, functional convergence or principles of folding? Q Rev Biophys 1980;13:339–386.
13. Murphy ME, Lindley PF, Adman ET. Structural comparison of cupredoxin domains: domain recycling to construct proteins with novel functions. Protein Sci 1997;6:761–770.
14. Ponting CP, Russell RB. Identification of distant homologues of fibroblast growth factors suggests a common ancestor for all beta-trefoil proteins. J Mol Biol 2000;302:1041–1047.
15. Copley RR, Bork P. Homology among (betaalpha)(8) barrels: implications for the evolution of metabolic pathways. J Mol Biol 2000;303:627–641.
16. Arcus V. OB-fold domains: a snapshot of the evolution of sequence, structure and function. Curr Opin Struct Biol 2002;12:794–801.
17. Kinch LN, Grishin NV. Evolution of protein structures and functions. Curr Opin Struct Biol 2002;12:400–408.
18. Panchenko AR, Madej T. Analysis of protein homology by assessing the (dis)similarity in protein loop regions. Proteins 2004;57:539–547.
19. Flores TP, Orengo CA, Moss DS, Thornton JM. Comparison of conformational characteristics in structurally similar protein pairs. Protein Sci 1993;2:1811–1826.
20. Wood TC, Pearson WR. Evolution of protein sequences and structures. J Mol Biol 1999;291:977–995.
21. Koehl P, Levitt M. Sequence variations within protein families are linearly related to structural variations. J Mol Biol 2002;323:551–562.
22. Marchler-Bauer A, Anderson JB, DeWeese-Scott C, Fedorova ND, Geer LY, He S, Hurwitz DI, Jackson JD, Jacobs AR, Lanczycki CJ, Liebert CA, Liu C, Madej T, Marchler GH, Mazumder R, Nikolskaya AN, Panchenko AR, Rao BS, Shoemaker BA, Simonyan V, Song JS, Thiessen PA, Vasudevan S, Wang Y, Yamashita RA, Yin JJ, Bryant SH. CDD: a curated Entrez database of conserved domain alignments. Nucleic Acids Res 2003;31:383–387.
23. Marchler-Bauer A, Panchenko AR, Shoemaker BA, Thiessen PA, Geer LY, Bryant SH. CDD: a database of conserved domain alignments with links to domain three-dimensional structure. Nucleic Acids Res 2002;30:281–283.
24. Berman HM, Bhat TN, Bourne PE, Feng Z, Gilliland G, Weissig H, Westbrook J. The Protein Data Bank and the challenge of structural genomics. Nat Struct Biol 2000;7(Suppl):957–959.
25. Gibrat JF, Madej T, Bryant SH. Surprising similarities in structure comparison. Curr Opin Struct Biol 1996;6:377–385.
26. Chen J, Anderson JB, DeWeese-Scott C, Fedorova ND, Geer LY, He S, Hurwitz DI, Jackson JD, Jacobs AR, Lanczycki CJ, Liebert CA, Liu C, Madej T, Marchler-Bauer A, Marchler GH, Mazumder R, Nikolskaya AN, Rao BS, Panchenko AR, Shoemaker BA, Simonyan V, Song JS, Thiessen PA, Vasudevan S, Wang Y, Yamashita RA, Yin JJ, Bryant SH. MMDB: Entrez's 3D-structure database. Nucleic Acids Res 2003;31:474–477.
27. Geer LY, Domrachev M, Lipman DJ, Bryant SH. CDART: protein homology by domain architecture. Genome Res 2002;12:1619–1623.
28. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–410.
29. Marchler-Bauer A, Bryant SH. Measures of threading specificity and accuracy. Proteins 1997;Suppl 1:74–82.
30. McLachlan AD. Gene duplications in the structural evolution of chymotrypsin. J Mol Biol 1979;128:49–79.
31. Preparata FP, Shamos MI. Computational geometry, an introduction. New York: Springer-Verlag; 1985.
32. Sokal RR, Rohlf FJ. Biometry. The principles and practice of statistics in biological research. New York: Freeman & Company; 1995.
33. Plaxco KW, Simons KT, Baker D. Contact order, transition state placement and the refolding rates of single domain proteins. J Mol Biol 1998;277:985–994.
34. Panchenko AR, Madej T. Structural similarity of loops in protein families: toward the understanding of protein evolution. BMC Evol Biol 2005;5:10–15.
35. Chothia C. Proteins. One thousand families for the molecular biologist [news]. Nature 1992;357:543–544.
36. Wolf YI, Grishin NV, Koonin EV. Estimating the number of protein folds and families from complete genome data. J Mol Biol 2000;299:897–905.
37. Coulson AF, Moult J. A unifold, mesofold, and superfold model of protein fold use. Proteins 2002;46:61–71.
38. Liu X, Fan K, Wang W. The number of protein folds and their distribution over families in nature. Proteins 2004;54:491–499.
39. Grant A, Lee D, Orengo C. Progress towards mapping the universe of protein folds. Genome Biol 2004;5:107.
40. Zhang C, DeLisi C. Estimating the number of protein folds. J Mol Biol 1998;284:1301–1305.
41. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
42. Yang AS, Honig B. An integrated approach to the analysis and modeling of protein sequences and structures. II. On the relationship between sequence and structural similarity for proteins that are not obviously related in sequence. J Mol Biol 2000;301:679–689.